DOCUMENT RESUME

ED 193 298 TM 800 613

AUTHOR Echternacht, Gary
TITLE No Treatment Expectations: Estimating the Unobserved and Implications for Interpreting NCE Gains.
PUB DATE Apr 80
NOTE 13p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Assessment: Elementary Secondary Education: *Evaluation Methods: *Program Evaluation: State School District Relationship: *Test Interpretation: Test Norms: Test Results: *Test Validity
IDENTIFIERS Elementary Secondary Education Act Title I: *Normal Curve Equivalent Scores: *Title I Evaluation and Reporting System

ABSTRACT

The Normal Curve Equivalent (NCE) gain statistic is examined, and considerations for its interpretation are highlighted. The NCE gain is made up of an observed and an expected part. The observed score is the posttest result. The expected score can never be observed nor verified for any of the Title I Evaluation and Reporting System (TIERS) models; it depends on a theory of growth to estimate expected performance, and each of the TIERS models assumes a different theory of growth. Therefore, NCE gains will vary according to the methodology employed. Five distinct considerations educators need to keep in mind when trying to interpret NCE gains include: (1) no matter what the NCE gain turns out to be, it could have come out differently: (2) evaluation results are valid to the extent that the achievement tests used are valid: (3) different methodologies will yield different results: (4) individual NCE gains must be treated with extreme caution: and (5) group results become more valid as the size of the group increases. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED193298

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

No Treatment Expectations: Estimating the Unobserved and Implications for Interpreting NCE Gains

by
Gary Echternacht
Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
G. Echternacht

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 800613

Presented at
The American Educational Research Association's
Annual Meeting
April 10, 1980

No treatment expectations: Estimating the unobserved and implications for interpreting NCE gains

Gary Echternacht
Educational Testing Service

Last fall ESEA Title I regulations were completed requiring school districts to report on the effectiveness of their projects by using one of three evaluation models. In the regulations they are listed as the norm referenced, control group, and special regression models. More commonly they are referred to as models A, B, and C. (See Tallmadge and Wood, 1976) Each evaluation model provides a measure of the effectiveness of the project on a metric that is common to all the models and is called the NCE scale. Results of local school district evaluations are aggregated at the state level and sent to USOE where they are aggregated with gains from other states to provide an indicator of how well ESEA Title I is improving the achievement levels of the students it serves.

Each of the evaluation models attempts to measure the effectiveness of a project in the same way. Observed levels of achievement at the end of the project are compared with estimates of how well students would have achieved at the same time without the compensatory program. This impact measure is often written as:

$$\text{Impact} = \text{Average (observed} - \text{expected)}$$

The evaluation models provide a means for obtaining the expected levels of performance. The special regression model (model C) and the control group model (model B) use test data from non compensatory program students in the local school district to derive a level of achievement

one might expect without Title I. These two models use methodologies that are often employed in social and psychological research and have been widely discussed in the measurement and statistical literature under the analysis of covariance and non-equivalent control group studies (see Kenny, 197, Rubin, 1977 for example). The norm referenced model (model A) uses test publisher norms to generate an expected level of performance. This methodology is not often discussed in the literature although variations of the norm referenced model have been used by local evaluators for some time.

Observed levels of performance in the impact estimate are simply observed posttest scores. Expected posttest scores are subtracted from the observed posttest scores and averaged for the students in Title I. Either observed and expected scores are expressed in NCES or the resulting difference is converted to an NCE gain. The NCE scale is a normalized standard score scale with a mean of 50 and a standard deviation of about 21. Positive NCEA gains indicate that achievement levels have improved over what would have been expected without the program.

The purpose of this paper is to take a close look at the NCE gain statistic and highlight what educators need to keep in mind when trying to understand, explain, and use NCE gain statistics. This is particularly important given the current thinking in government about the utility of local district project evaluation. There are strong forces encouraging local districts to use the evaluation results to improve programs (see Barnes and Ginsburg, 1979). One reason that local evaluation has been stressed over large scale national evaluations is the belief that local

evaluation is beneficial to local program improvement while large scale national evaluations are not. The USOE regional technical assistance center project is developing and giving numerous workshops on how to use evaluation data. NIE has recently contracted with the Huron Institute to study exemplary ways in which evaluation data can be used. State departments of education want to assess the effectiveness of the projects they fund. Indeed, this entire symposium is a testament to the notion that the resulting NCE gains should be and will be used to change programs in some way.

Disecting the impact estimate

As indicated earlier, the NCE gain is made up of an observed and expected part. The observed score is simply the posttest result which is generally obtained one or two months before the end of the spring term. At issue with the posttest are only the reliability and validity of the test used. If the test is one of the commercially available norm referenced tests, reliability is not much of an issue. Validity, or what the test is supposed to measure, may be an issue. Generally speaking, tests used for program evaluation should measure the content areas included in the Title I curriculum. In some school districts choice of the tests used for program evaluation are dictated by people who have little knowledge of what is being taught in Title I. In other school districts, people attempt to economize on testing by using one test for a multitude of purposes--e.g., for placement, selection, evaluation, diagnosis. One test cannot serve these purposes equally well. If a test is chosen primarily because it is a good diagnostic test, it may be lacking somewhat as a valid evaluation test.

As an example of how using a test for multiple purposes affects gains, let me give you the case of the Stanford Diagnostic Reading Test. This test was designed to provide teachers with diagnostic information about low ability students in reading. It is an excellently constructed test that provides much useful information to teachers. It also has national norms. These norms are cross sectional norms. The same students who took the test in the fall norming are not the same students who took the test in the spring norming. In the norming, the test was not used in a diagnostic-prescriptive way. If the Stanford Diagnostic were used in Title I both as a diagnostic-prescriptive tool and as an evaluation tool, extremely high gains would seem likely in a model A1 design. If I were an evaluator in a state department of education, I would closely monitor results by test to see if the above is true.

Regulations for Title I evaluation require that local districts evaluate their tests with respect to reliability and validity. Unlike the evaluation models, specific procedures for assessing reliability and validity have not been spelled out. How this requirement will be monitored by USOE and the states has also yet to be established. We can, however, be assured that if the tests used for evaluation are not valid, then neither can the resulting NCE gains be valid.

The second part of the impact estimate is much more interesting than the simple posttest score. It represents a level of performance that would have been observed had the student not received Title I instruction. But the student did receive Title I instruction. That is the evaluation design problem. The expected level of performance can never be observed.

It can never be verified for any of the models. It is a theoretical construct, and we must use a theory of growth to estimate expected performance. But there is an additional problem. Theories of growth differ. Each of the evaluation models assumes a different theory of growth, as do some variations of specific models. Different theories produce different expectations, meaning that NCE gains will vary according to the methodology employed. This means that although two school districts using different methodologies both have NCE gains, those gains are not strictly comparable. Any such comparison introduces some as yet unknown amount of bias and error.

When we look closer at how expected levels of performance are estimated, we find that they are based on the performance of a group not receiving compensatory instruction. As the composition of that group changes so too will the expectancy. If schools not having Title I projects are added to a model C analysis, the expectancy will change. If different achievement tests are used in model A, the expectancy will change. If a Title I group is highly mobil and the number of students with both pretest and posttest scores is small, the expectancy will change in model A depending on just who migrates and who does not.

Expected levels of performance are also closely dependent on the way in which they are interpreted. Let us assume that there were no Title I or other compensatory programs and that we administered a test nationally to every student in a certain grade at two times. If we were to plot those scores, they might appear as in Figure 1. If all scores were in terms of NCEs, we could draw a principal axis, or in this case a

45 degree line showing where pretest and posttest scores are equal. We could also draw the regression of posttest or pretest.

Insert Figure 1
about here

If we look at individuals who score x_o on the pretest, models A and B will lead us to conclude that the expected level of performance is y_1. Model C will lead us to conclude that y_2 is the expected level of performance. Mary's level of performance is y_o which is neither y_1 nor y_2. To complicate things more, y_o is the most valid expected level of performance for Mary because y_o was her observed level of performance in the absence of a compensatory program. The expectancies, y_1 and y_2 or conditional expectancies, obtained from using the models represent typical expectancies for all those scoring x_o on the pretest even though those scoring x_o on the pretest display a wide range of posttest scores.

Thus, we can see that the expectancies resulting from using the evaluation models are those typical of people scoring at a specific x_o on the pretest. Individual NCE gains are going to be more subject to error than we would expect because another source of variation--variation among individuals at the same pretest level--must be included in the standard error. Individual NCE gains must be interpreted with much more care than classroom, building, or school district gains.

In most instances, NCE gains will be used to judge the performance of groups. As we saw with the relationship between individual and conditional expectancies, as the size of the group for which we obtain an expectancy increases, the variation in expectancies decreases. Thus, building expectancies will be less variable than will classroom expectancies. School district expectancies will be less variable than will building expectancies.

Expectancies, NCE gains, and their use

I hope that I have been able to clearly convey five distinct messages about interpreting NCE gains. Let me list them:

1. No matter what the NCE gain turns out to be, it could have come out differently.

Different methodologies, different tests, a different sample of students all mean different NCE gains. NCE gains from projects using different methodologies should not be compared. But even within methodologies, we must be careful. A responsible approach would be to ask, "just how different might we expect NCE gains to be?" When we ask this question, we approach another level of analysis. We move away from simple indication toward description. We begin to estimate standard errors for the gains and speak in terms of confidence levels. More care is placed in interpretation as we move away from the NCE gain as an absolute indicator of performance and toward the NCE gain as a general descriptor, a statistic that is used to trigger a more in depth study.

2. Evaluation results are valid to the extent that the achievement tests used are valid.

School districts are looking for ways to economize on their testing. At the same time, school districts are being asked to use test results in more ways. Asking a single test to provide information for selection, placement and evaluation necessarily results in increased inefficiency. If a test is chosen because it provides good diagnostic information, it will not be as valid as a test chosen specifically for evaluation. I would suggest that school districts who use their evaluation tests primarily for some other purpose interpret those results with great caution. State departments of education should also keep this in mind when interpreting NCE gains.

3. Different methodologies will yield different results.

Each evaluation model estimates a level of achievement that would have been observed had there been no Title I program. Each model is based on a different theory of achievement growth. The results from such evaluations are bound to be different. Districts using a prescribed methodology should only compare their results with other districts using a similar methodology.

4. Individual NCE gains must be treated with extreme caution.

Individual gains are much more varied than are group gains. When we use an evaluation model to obtain an expected level of performance, we are using group data. These expectations are not standards but rather the typical performance of students with the same pretest score. Individual expectations will vary around this group expectation, adding another source of variation to an individual expectation.

5. Group results become more valid as the size of the group increases.

Elementary statistics teaches us that as sample size increases, the accuracy of the estimate of the mean from that population increases. The Title I reporting system does not require that standard errors or confidence limits be set. That is unfortunate because by not obtaining standard errors an unwarranted authority is ascribed to the precision of an NCE gain for small samples.

The implication of the above messages is that neither states nor local school districts should take a single NCE gain statistic too seriously. As we begin to accumulate NCE gain over a number of years, however, we can place more credence in the results. For example, if I were a Title I director and achieved a very low NCE gain, I would dismiss the statistic as being insufficiently reliable to make other than minor changes. If the results were repeated a second year, I would start to look at my project management, curriculum materials, and evaluation methodology. If the results were repeated a third year, I would believe that I had a definite problem and would begin an analysis of the program, its promises, and its premises.

Another implication of the above is we need to know much more about the non-programmatic factors--e.g., tests, designs, time frames--before we can interpret NCE gains. State would be wise to make analyses of the gains their districts obtain. The results of these analyses would greatly aid in interpreting NCE gains and in using these gains as either a trigger to identify districts with problems or districts that perform well above the norm.

References

Barnes, R. and Ginsburg, A. Relevance of the RMC models for Title I policy concerns. Educational Evaluation and Policy Analysis. 1979, 1, (2), p. 7-14.

Kenny, D. A quasi-experimental approach to assessing treatment effects in nonequivalent control group designs. Psychological Bulletin, 1979, 82 (3), 345-362.

Rubin, D. Assignment to treatment group on the basis of a covariate. Journal of Educational Statistics, 1977, 2, (1), 1-26.

Tallmadge, G. and Wood, C. User's Guide: ESEA Title I evaluation and reporting system. RMC Corporation: Mountain View, CA, 1976.

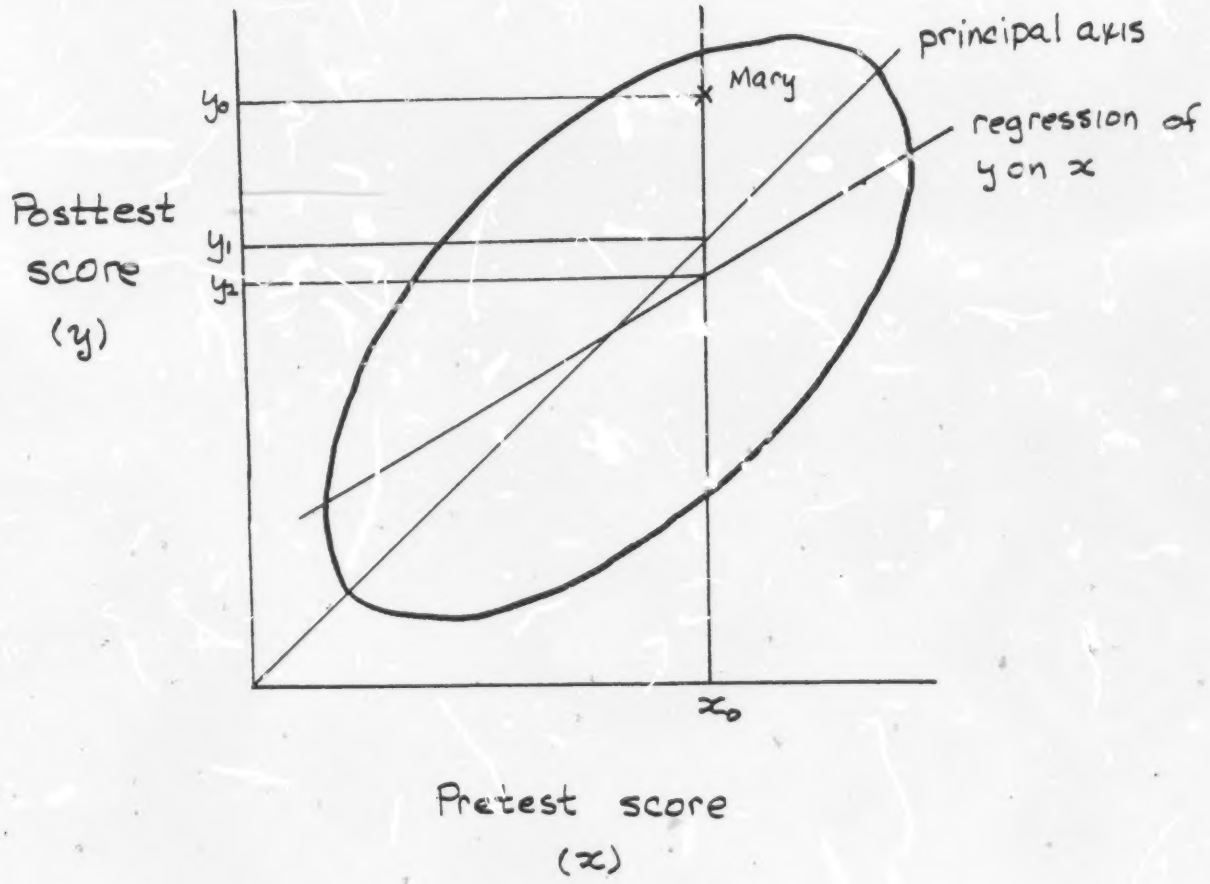

Figure 1. No treatment performance between pretest and posttest.